Clearer Frames, Anytime: Resolving Velocity Ambiguity in Video Frame Interpolation

Zhihang Zhong[1], Gurunandan Krishnan[2], Xiao Sun[1], Yu Qiao[1],
Sizhuo Ma[2], Jian Wang[2]
[1]Shanghai Artificial Intelligence Laboratory, [2]Snap Inc.

ECCV 2024, Oral

Outline

- Introduction: **Video frame interpolation**
- Problem: **Velocity ambiguity in time indexing**
- Methodology: **Strategies for disambiguation**
- Experiment: **Effectiveness of plug-and-play strategies**
- New feature: **Manipulated interpolation of anything**
- Conclusion and future work

Introduction: **Video frame interpolation**

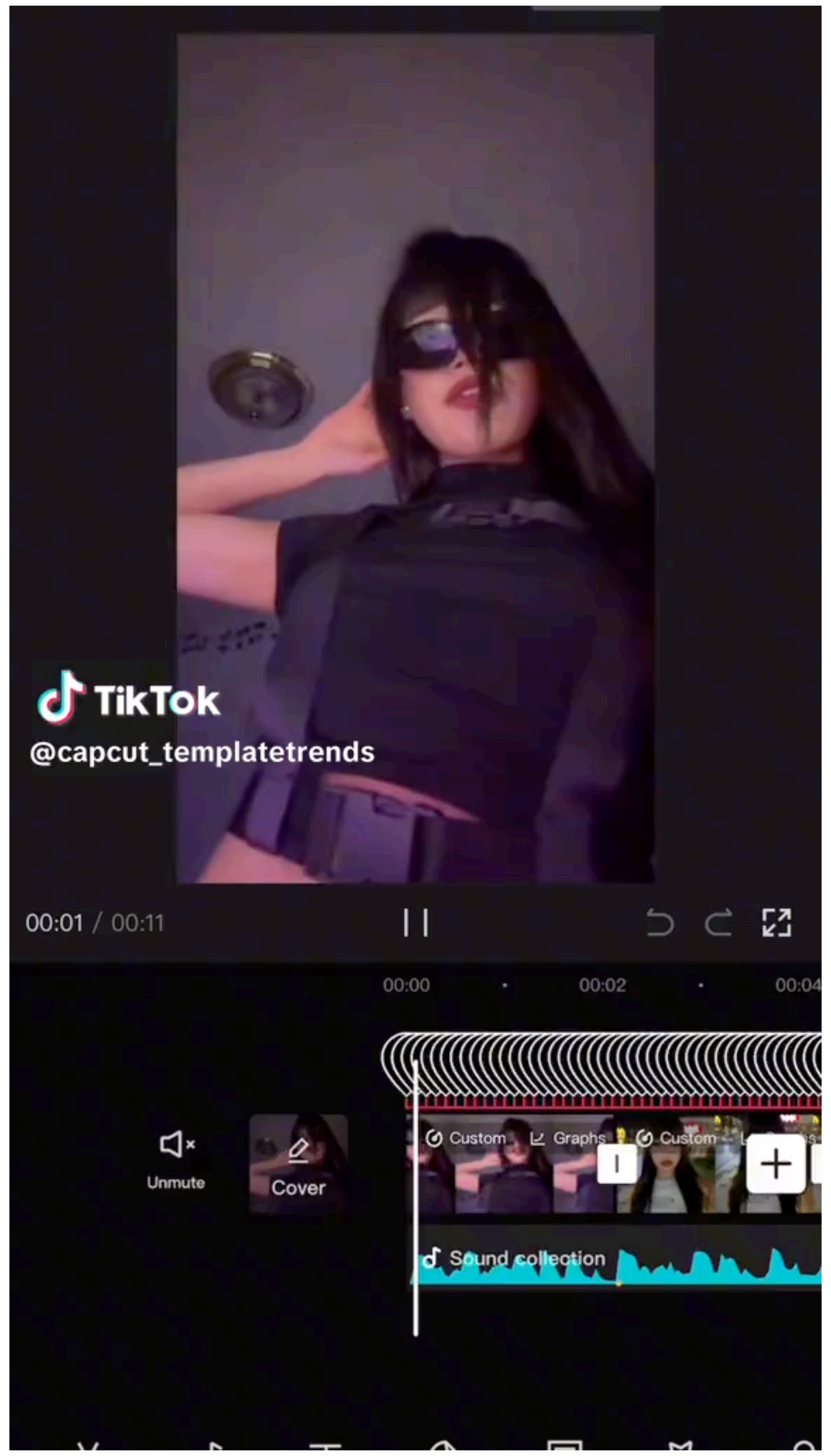
- Video frame interpolation (VFI) has wide applications

Slow motion of highlights

Original Videos from Legendary Shots Youtube Channel

Original Videos from Legendary Shots Youtube Channel

Sync video to the beat



Assisting video generation

Video compression

H.264
MPEG-4/AVC

Introduction: **Paradigms**

- Traditional flow-based approaches: **linear motion; holes**
- Learning-based approaches include **fixed-time** & **arbitrary-time** interpolation
- Arbitrary-time: **faster for any timestep; no accumulation errors**

$$I_t = \mathcal{F}\left(I_0, I_1, t\right).$$

I_0

I_1

arbitrary time $t \in [0, 1]$

Problem: **Velocity ambiguity in time indexing**

- The velocities of individual objects within starting and ending frames remain undefined, introducing a velocity ambiguity, a myriad of plausible time-to-location mappings during training

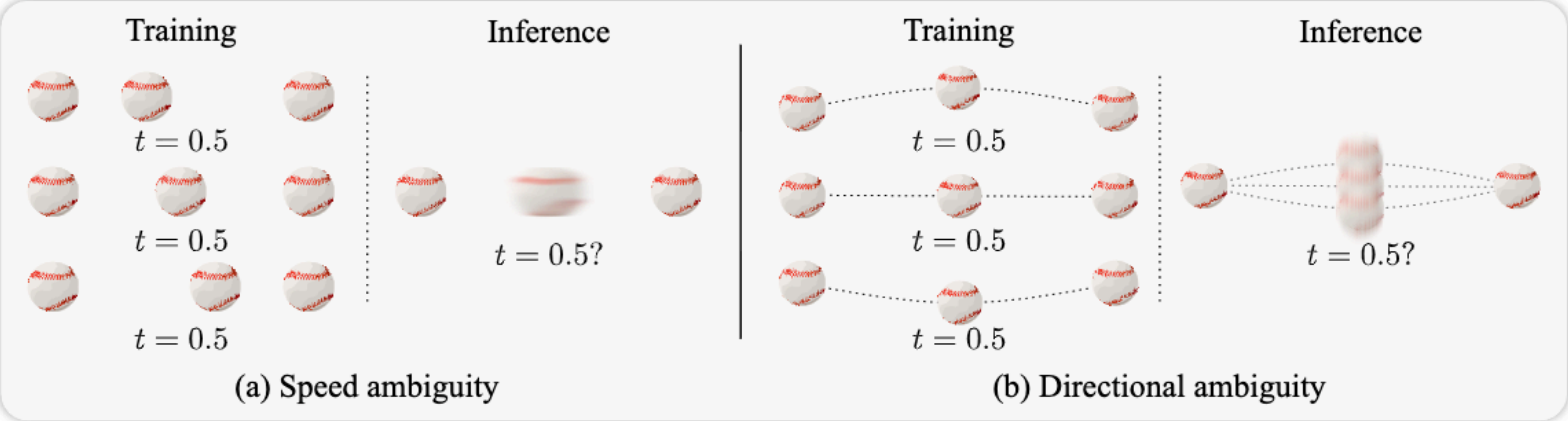
$$\{I_t^1, I_t^2, \ldots, I_t^n\} = \mathcal{F}(I_0, I_1, t),$$

- As a result, models trained with ***time indexing*** tend to produce blurred and imprecise interpolations, as they average out the potential outcomes.

$$\hat{I}_t = \mathbb{E}_{I_t \sim \mathcal{F}(I_0, I_1, t)}[I_t].$$

Problem: **Velocity ambiguity in time indexing**

- Velocity ambiguity encompasses **speed ambiguity** & **directional ambiguity**



(a) Speed ambiguity

(b) Directional ambiguity

Methodology: **Strategies for disambiguation**

- ***Could an alternative indexing method minimize such conflicts?***

$$\boxed{I_t = \mathcal{F}\left(I_0, I_1, \text{motion hint}\right) \Rightarrow I_t = \mathcal{F}\left(I_0, I_1, D_t\right).}$$

- Optical flow? → Unknown at inference time

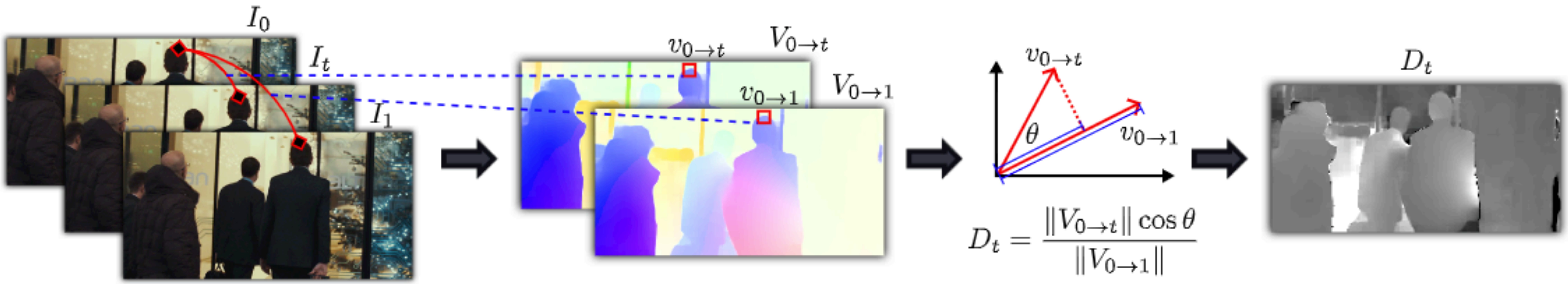
- Instead, we propose a more flexible ***distance indexing*** approach. We employ a ***distance ratio*** map D_t, where each pixel denotes ***how far the object has traveled between start and end frames***, within a normalized range of [0,1]

Methodology: **Distance indexing**

- We guide the network to interpolate more precisely without relying on the ambiguous time-to-location mapping to decipher it independently

$$I_t = \mathcal{F}(I_0, I_1, \mathcal{D}(t)). \quad \rightarrow \quad I_t = \mathcal{F}(I_0, I_1, D_t).$$

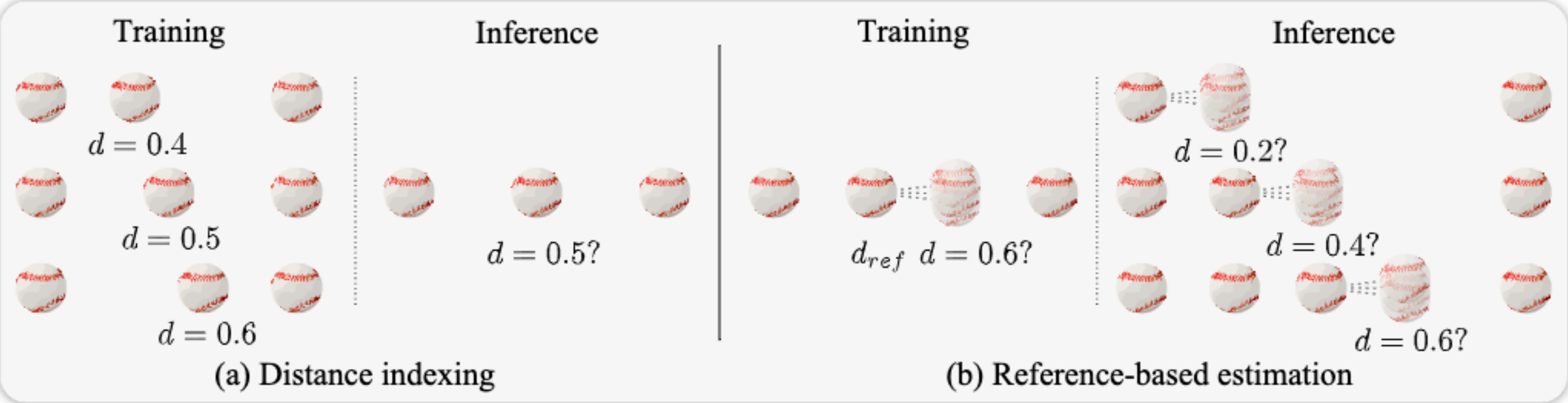
- ***In practice, we notice it is sufficient to provide a uniform map $D_t = t$, similar to time indexing (move each object at constant speeds along trajectories)***



$$D_t = \frac{\|V_{0\to t}\| \cos\theta}{\|V_{0\to 1}\|}$$

Methodology: **Iterative reference-based estimation**

- Although ***distance indexing (a)*** addresses the scalar ***speed ambiguity***, the ***directional ambiguity*** of motion remains a challenge.
- We introduce an ***iterative reference-based estimation strategy (b)***,which incrementally estimates distances, beginning with nearby points and advancing to farther ones, to mitigate the remained ambiguity

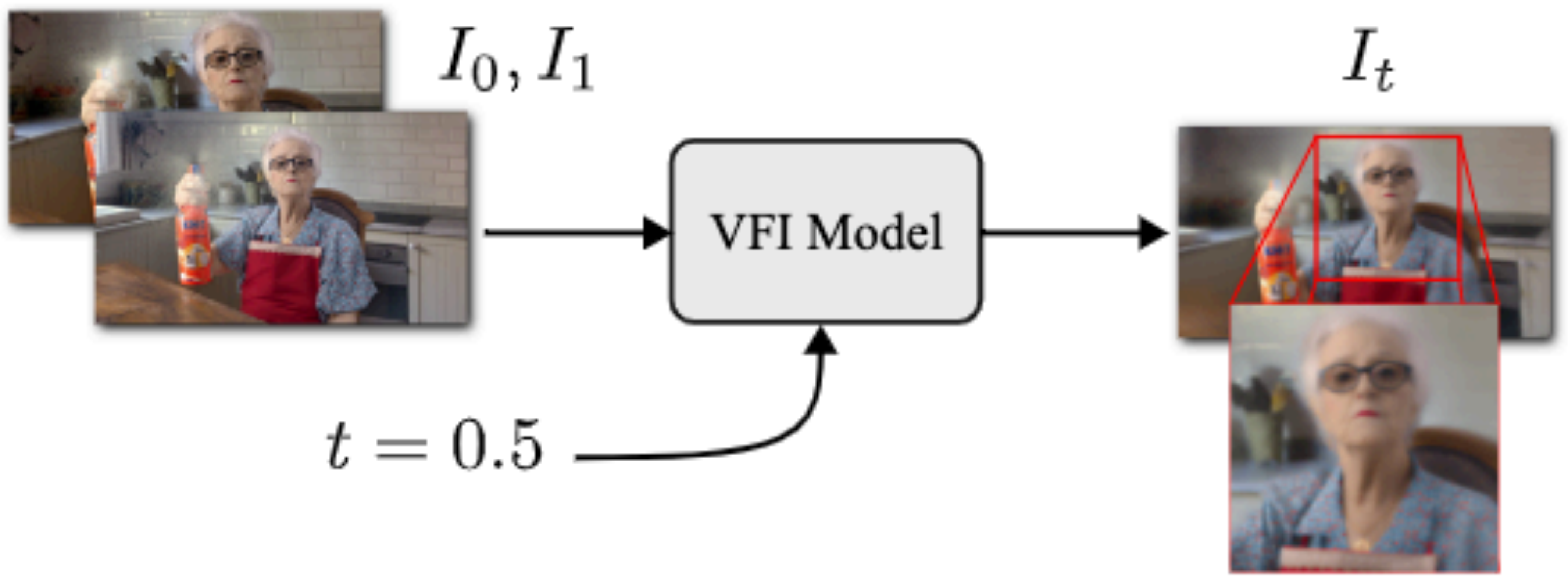


(a) Distance indexing

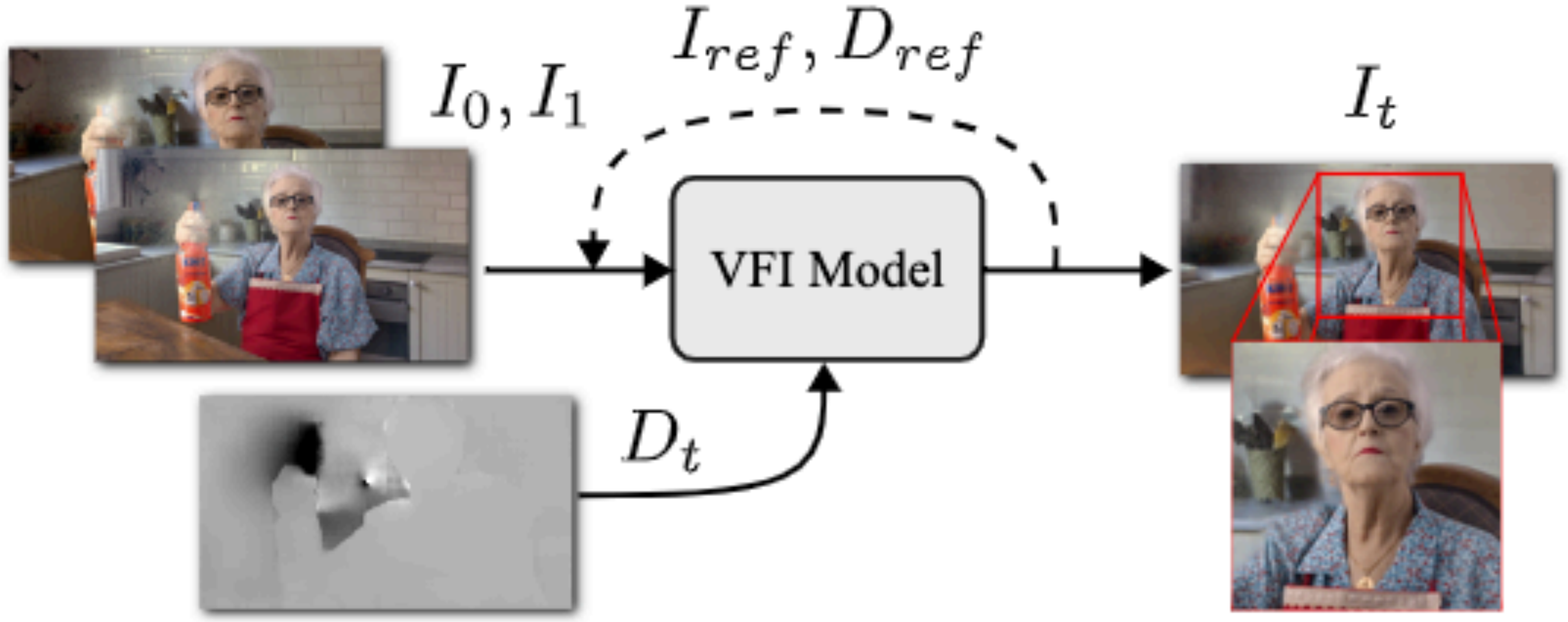
(b) Reference-based estimation

Methodology: **Plug-and-play**

- Our approach addresses challenges that are not bound to specific network architectures. Indeed, it can be applied as a plug-and-play strategy that requires only modifying the input channels for each model



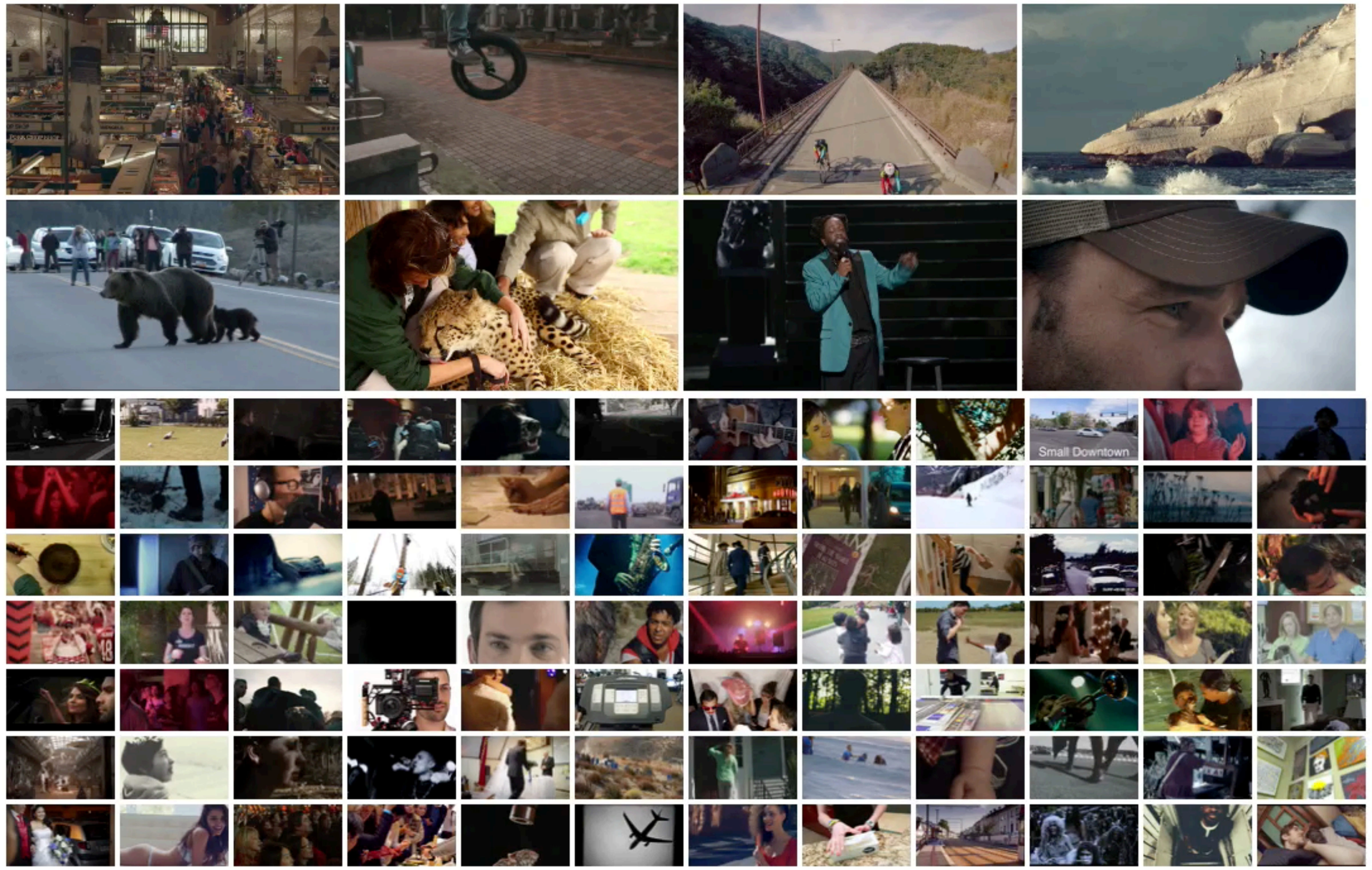
(a) Training paradigm of time indexing



(b) Training paradigm of distance indexing

Experiments: **Vimeo90K septuplet dataset**

- Consists of 91,701 seven frame sequences with fixed resolution 448 x 256, extracted from 39,000 selected video clips

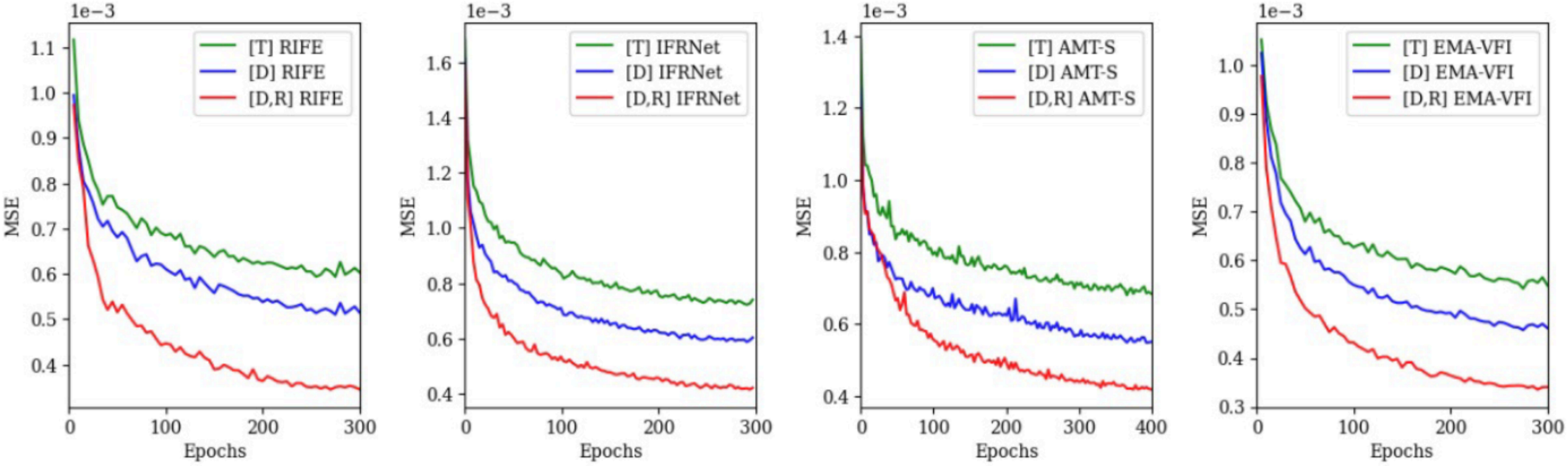


Experiment: **State-of-the-art models**

- [ECCV 2022] Real-Time Intermediate Flow Estimation for Video Frame Interpolation
- [CVPR 2022] IFRNet: Intermediate Feature Refine Network for Efficient Frame Interpolation
- [CVPR 2023] Extracting Motion and Appearance via Inter-Frame Attention for Efficient Video Frame Interpolation (EMA-VFI)
- [CVPR 2023] AMT: All-Pairs Multi-Field Transforms for Efficient Frame Interpolation
- **[T] time indexing; [D] distance indexing; [R] reference-based estimation**

Experiment: **Convergence curves**

- [D] and [R] facilitate the convergence of each model



(a) RIFE

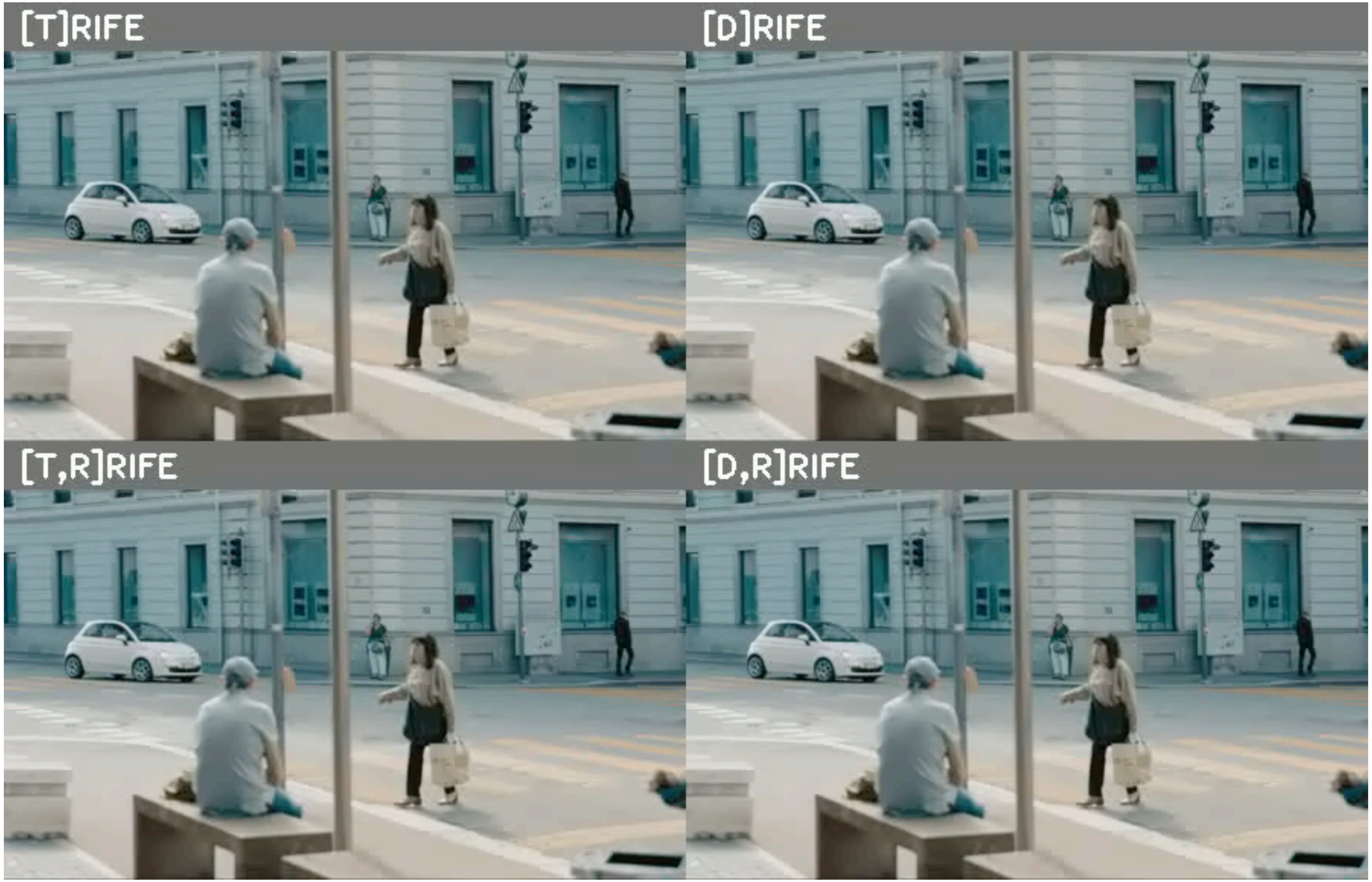
(b) IFRNet

(c) AMT

(d) EMA-VFI

Experiment: **Qualitative**

Experiment: **Qualitative**



Experiment: **Quantitative**

EUROPEAN CONFERENCE ON COMPUTER VISION

MILANO 2024

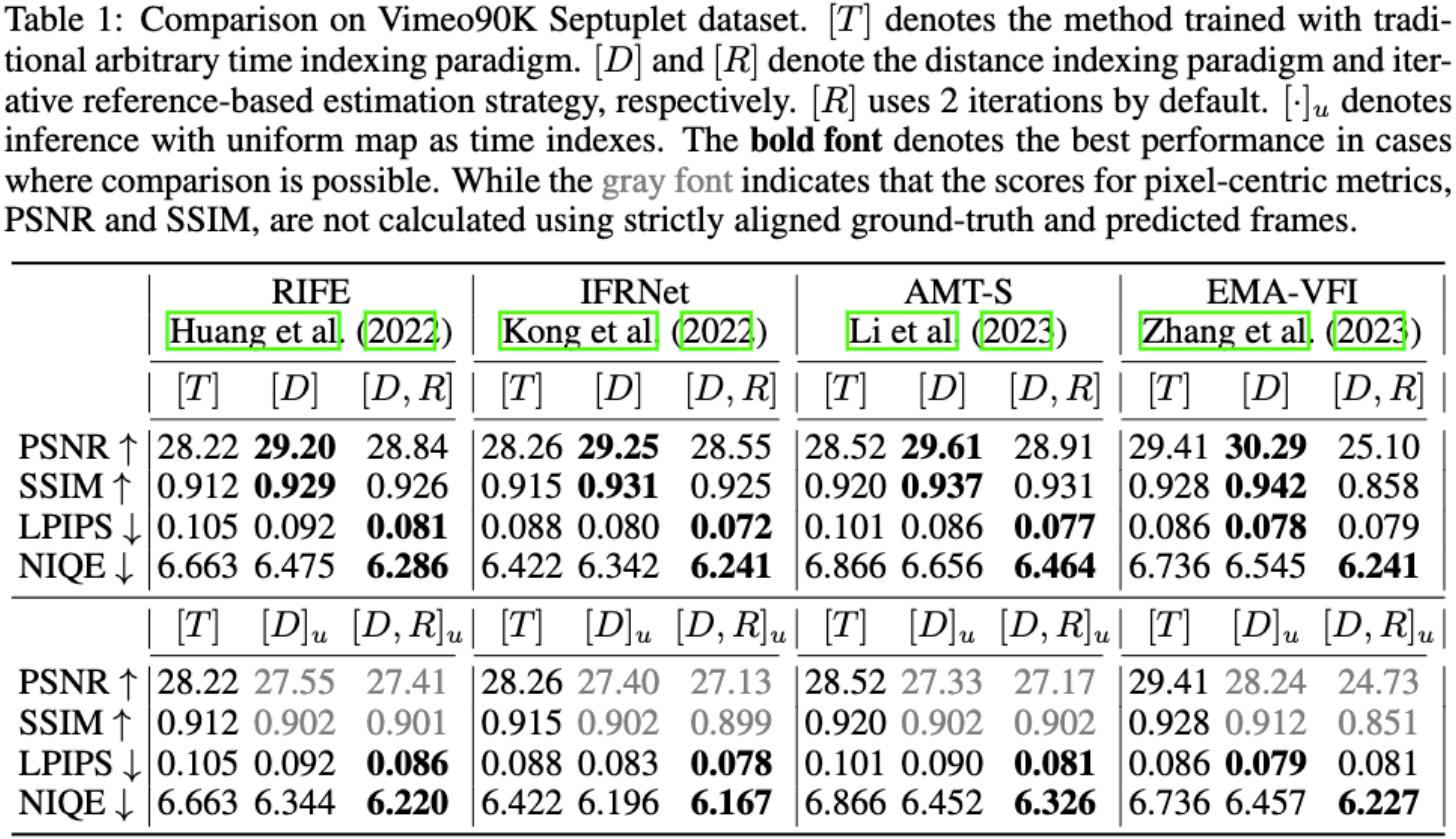


Table 1: Comparison on Vimeo90K Septuplet dataset. $[T]$ denotes the method trained with traditional arbitrary time indexing paradigm. $[D]$ and $[R]$ denote the distance indexing paradigm and iterative reference-based estimation strategy, respectively. $[R]$ uses 2 iterations by default. $[\cdot]_u$ denotes inference with uniform map as time indexes. The **bold font** denotes the best performance in cases where comparison is possible. While the gray font indicates that the scores for pixel-centric metrics, PSNR and SSIM, are not calculated using strictly aligned ground-truth and predicted frames.

	RIFE Huang et al. (2022)			IFRNet Kong et al. (2022)			AMT-S Li et al. (2023)			EMA-VFI Zhang et al. (2023)		
	$[T]$	$[D]$	$[D,R]$	$[T]$	$[D]$	$[D,R]$	$[T]$	$[D]$	$[D,R]$	$[T]$	$[D]$	$[D,R]$
PSNR ↑	28.22	**29.20**	28.84	28.26	**29.25**	28.55	28.52	**29.61**	28.91	29.41	**30.29**	25.10
SSIM ↑	0.912	**0.929**	0.926	0.915	**0.931**	0.925	0.920	**0.937**	0.931	0.928	**0.942**	0.858
LPIPS ↓	0.105	0.092	**0.081**	0.088	0.080	**0.072**	0.101	0.086	**0.077**	0.086	**0.078**	0.079
NIQE ↓	6.663	6.475	**6.286**	6.422	6.342	**6.241**	6.866	6.656	**6.464**	6.736	6.545	**6.241**
	$[T]$	$[D]_u$	$[D,R]_u$	$[T]$	$[D]_u$	$[D,R]_u$	$[T]$	$[D]_u$	$[D,R]_u$	$[T]$	$[D]_u$	$[D,R]_u$
PSNR ↑	28.22	27.55	27.41	28.26	27.40	27.13	28.52	27.33	27.17	29.41	28.24	24.73
SSIM ↑	0.912	0.902	0.901	0.915	0.902	0.899	0.920	0.902	0.902	0.928	0.912	0.851
LPIPS ↓	0.105	0.092	**0.086**	0.088	0.083	**0.078**	0.101	0.090	**0.081**	0.086	**0.079**	0.081
NIQE ↓	6.663	6.344	**6.220**	6.422	6.196	**6.167**	6.866	6.452	**6.326**	6.736	6.457	**6.227**

Experiment: **Quantitative**

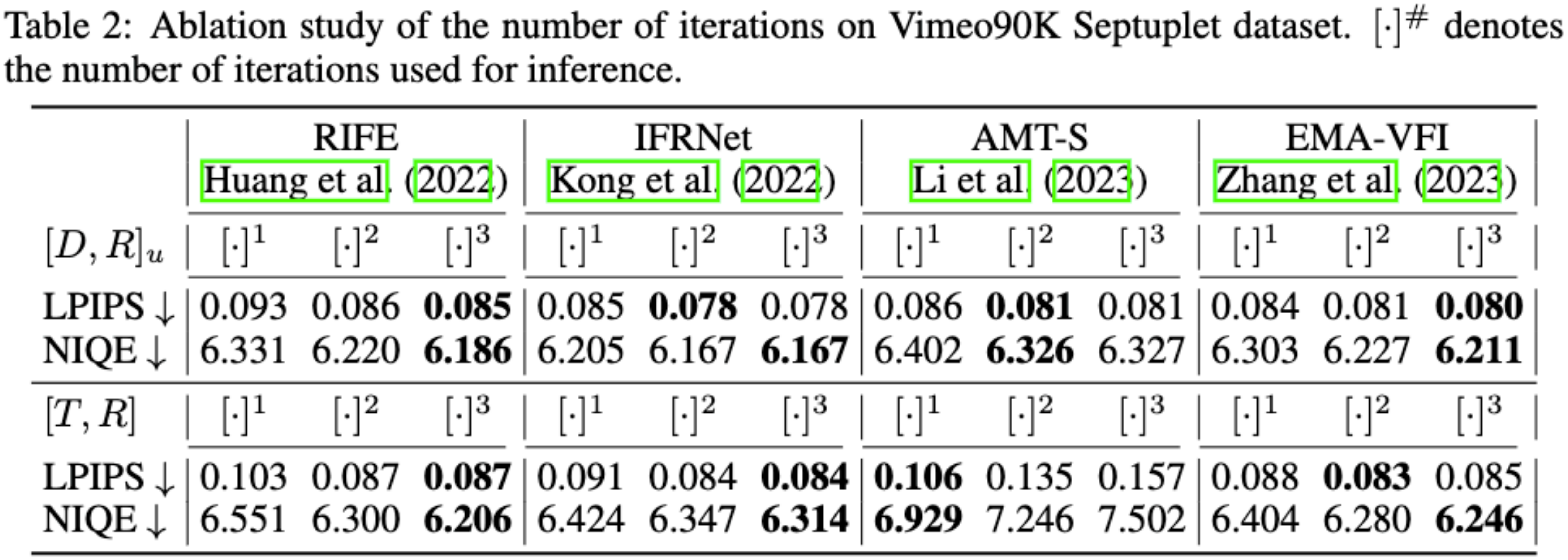
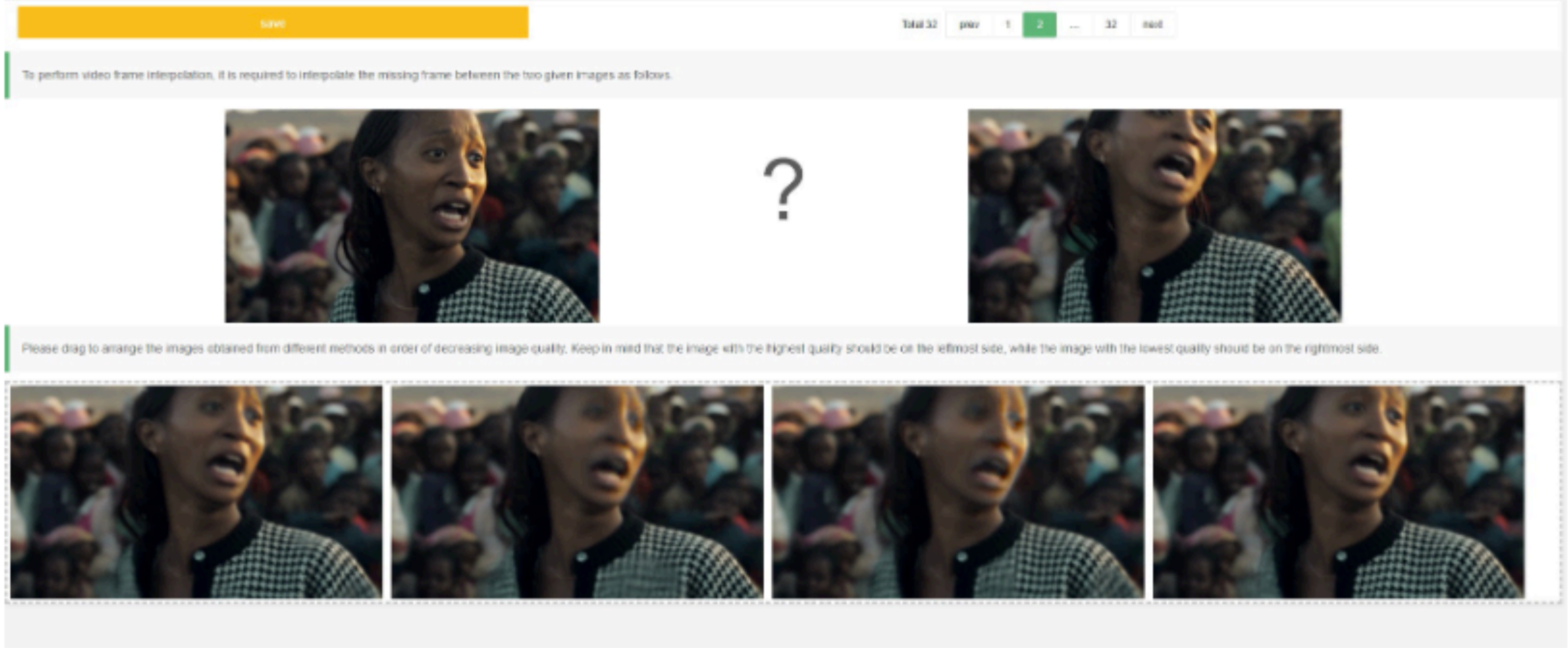


Table 2: Ablation study of the number of iterations on Vimeo90K Septuplet dataset. $[\cdot]^{\#}$ denotes the number of iterations used for inference.

	RIFE Huang et al. (2022)			IFRNet Kong et al. (2022)			AMT-S Li et al. (2023)			EMA-VFI Zhang et al. (2023)		
$[D, R]_u$	$[\cdot]^1$	$[\cdot]^2$	$[\cdot]^3$	$[\cdot]^1$	$[\cdot]^2$	$[\cdot]^3$	$[\cdot]^1$	$[\cdot]^2$	$[\cdot]^3$	$[\cdot]^1$	$[\cdot]^2$	$[\cdot]^3$
LPIPS ↓	0.093	0.086	**0.085**	0.085	**0.078**	0.078	0.086	**0.081**	0.081	0.084	0.081	**0.080**
NIQE ↓	6.331	6.220	**6.186**	6.205	6.167	**6.167**	6.402	**6.326**	6.327	6.303	6.227	**6.211**
$[T, R]$	$[\cdot]^1$	$[\cdot]^2$	$[\cdot]^3$	$[\cdot]^1$	$[\cdot]^2$	$[\cdot]^3$	$[\cdot]^1$	$[\cdot]^2$	$[\cdot]^3$	$[\cdot]^1$	$[\cdot]^2$	$[\cdot]^3$
LPIPS ↓	0.103	0.087	**0.087**	0.091	0.084	**0.084**	**0.106**	0.135	0.157	0.088	**0.083**	0.085
NIQE ↓	6.551	6.300	**6.206**	6.424	6.347	**6.314**	**6.929**	7.246	7.502	6.404	6.280	**6.246**

Experiment: **User study**

- Questionnaire statistics of VFI model's performance (Webapp)
- Ranking of [T], [D], [T,R], and [D,R]
- 30 anonymous participants

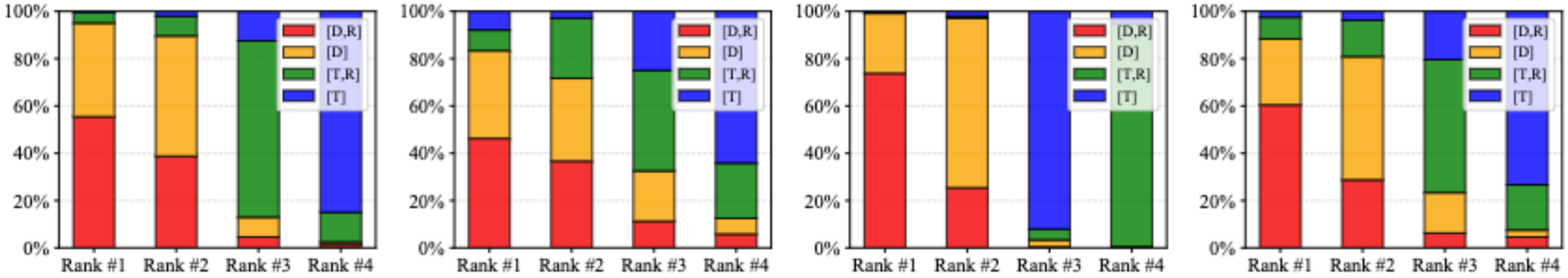


To perform video frame interpolation, it is required to interpolate the missing frame between the two given images as follows.

Please drag to arrange the images obtained from different methods in order of decreasing image quality. Keep in mind that the image with the highest quality should be on the leftmost side, while the image with the lowest quality should be on the rightmost side.

Experiment: **User study**

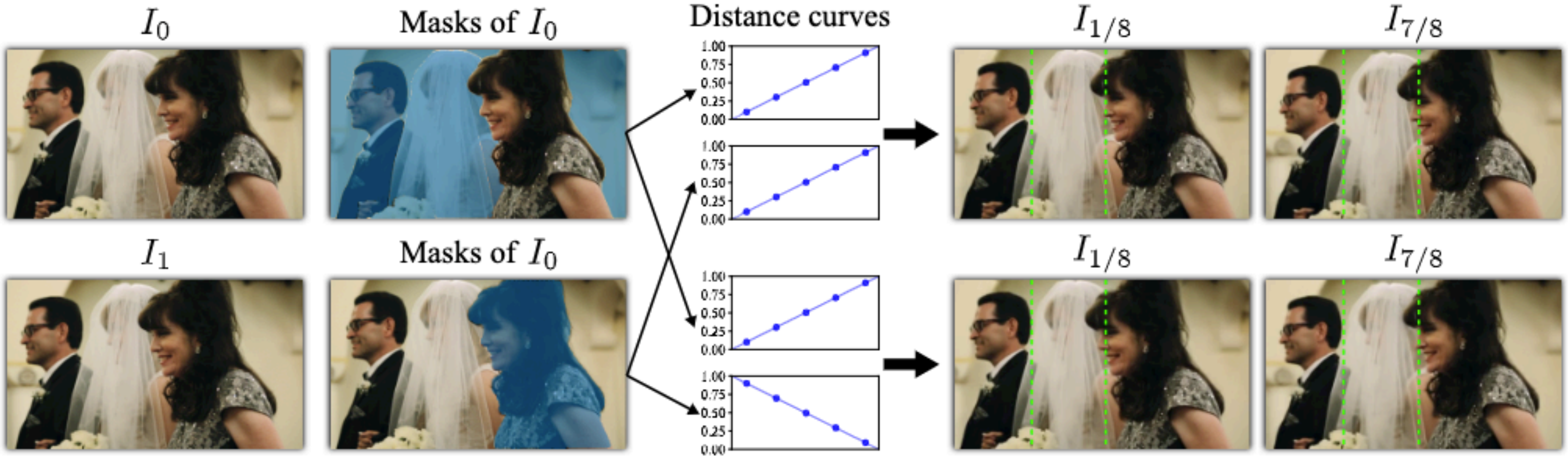
- The results align with our qualitative and quantitative findings. The [D,R] model variant emerged as the top-rated, underscoring the effectiveness of our strategies



(a) RIFE (b) IFRNet (c) AMT-S (d) EMA-VFI

New Feature: **Manipulated interpolation of anything**

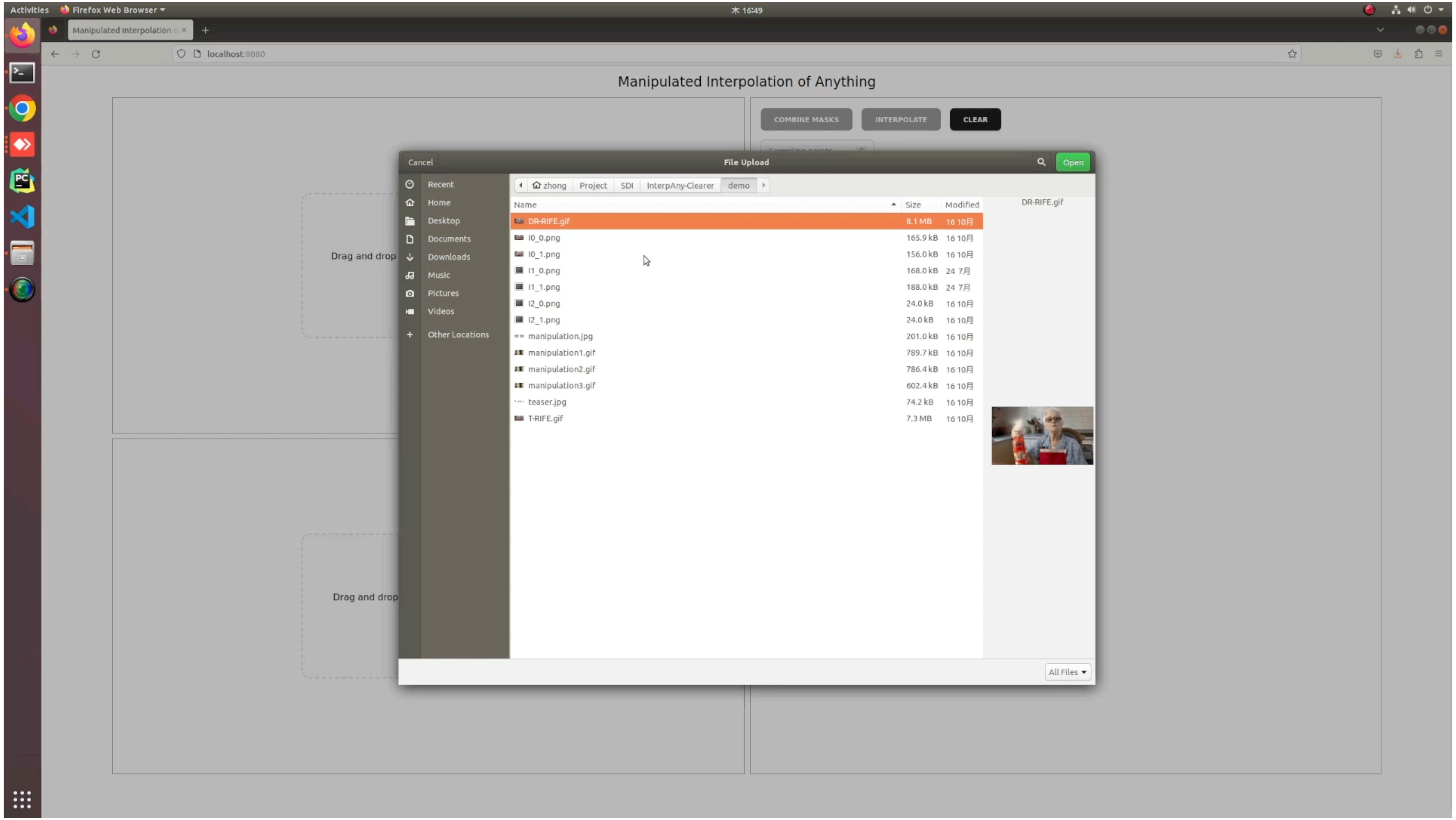
- Instead of using a uniform map, it is also possible to use a spatially-varying 2D map as input to manipulate the motion of objects. Paired with SOTA segmentation models such as SAM, this empowers users to freely control the interpolation of any object, *e.g.*, making certain objects backtrack in time



New Feature: **Demo of webapp**

EUROPEAN CONFERENCE ON COMPUTER VISION

MILANO 2024



Name	Size	Modified
DR-RIFE.gif	8.1 MB	16 10月
I0_0.png	165.9 kB	16 10月
I0_1.png	156.0 kB	16 10月
I1_0.png	168.0 kB	24 7月
I1_1.png	188.0 kB	24 7月
I2_0.png	24.0 kB	16 10月
I2_1.png	24.0 kB	16 10月
manipulation.jpg	201.0 kB	16 10月
manipulation1.gif	789.7 kB	16 10月
manipulation2.gif	786.4 kB	16 10月
manipulation3.gif	602.4 kB	16 10月
teaser.jpg	74.2 kB	16 10月
T-RIFE.gif	7.3 MB	16 10月

Conclusion and future work

- We propose distance indexing and iterative reference-based estimation to address the velocity ambiguity and enhance the capabilities of arbitrary time interpolation models

- We present an unprecedented manipulation method that allows for customized interpolation of any object

- Using multiple frames to estimate an accurate distance ratio map for a specific object is one of future works